Chapter 27: Measuring L2 speaking

Fumiyo Nakatsuhara, Chihiro Inoue and Nahal Khabbazbashi

Key Concepts

Levelt's (1989) model of speaking: This model assumes that separate cognitive processes are responsible for different aspects of speech production. It comprises three main systems: Conceptualizer, Formulator and Articulator. For L2 learners, incomplete L2 knowledge and lack of automaticity are thought to affect the Formulator and Articulator, resulting in less accurate and fluent speech. This model is key when we explore the cognitive validity of speaking tasks (Field, 2011).

Processing competence: The operational definition of processing competence is efficiency of processing language (Van Moere, 2012). It is thought to be a stable and measurable construct that underlies language proficiency, and automated speaking assessment has great potential for tapping into aspects of this construct.

Interactional competence: It is "the ability to co-construct interaction in a purposeful and meaningful way, taking into account sociocultural and pragmatic dimensions of the speech situation and event" (Galaczi & Taylor, 2018: 226), which is most appropriately assessed through paired and group oral discussion formats.

Background

Most language educators would agree that direct assessment of speaking is a resource-intensive, logistically-complex endevour and have seen a pragmatic decision made to compromise the ways in which speaking is assessed due to practical constraints in resource-limited contexts (Vidaković & Galaczi, 2013). The status of speaking ability in second language (L2) teaching and learning can however be traced back to the late 19th century in Europe which saw the rise of the *direct method* and the *oral method* with a primary emphasis on oral fluency (e.g. Palmer, 1921) and the *Reform Movement* in language education (e.g. Sweet, 1899) which stressed the importance of spoken language with particular focus on phonetics. The core role of speaking in L2 teaching was reflected in the testing approach adopted in the Cambridge Certificate of Proficiency in English (CPE) launched in 1913. The original CPE test (1913-45) assessed knowledge of phonetics in addition to a 1-hour speaking component which involved dictation, read-aloud and conversation tasks in an examiner-candidate direct speaking format. The test initially targeted English teacher trainees, and is one of the earliest examples of a speaking construct included in L2 proficiency tests (Vidaković & Galaczi, 2013).

What paved the way for a dramatic shift in L2 speaking assessment practices in the mid-20th century was the experience of the World Wars which highlighted a shortage of military personnel proficient in speaking a foreign language. In the US, the increased interest in enhancing speaking skills led to the development of the US Foreign Service Institute (FSI) Oral Proficiency Interview (OPI) introduced in 1952. The OPI test was noteworthy in the history of oral assessment on two accounts: firstly, its original holistic scale included performance descriptors defining features of distinct levels of proficiency and secondly, its revised scale incorporated analytic components (accent, comprehension, fluency, grammar and vocabulary). Therefore, it made a significant step towards defining a multi-faceted speaking construct and to assessing it reliably (Fulcher, 2003). The FSI OPI test and its rating scale gave rise to subsequent oral assessment initiatives in the US in the 1960s to 1980s, including the Interagency Language Roundtable speaking test, and the American Council for the Teaching of Foreign Languages (ACTFL) OPI. In the UK, it was also during the 1970s when the University of Cambridge Local Examinations Syndicate (UCLES) started denoting performance descriptors to standardise ratings by their oral examiners. As such, "the 1960s to 1980s … saw a growth in the explicit description of performance in assessment scales, and by extension, in the explicit definition of the test construct via the scales" (Vidaković & Galaczi, 2013: 268).

It should be noted that the rating scales and performance descriptors developed at the time were not informed by empirical research but were rather based on expert intuition (Fulcher, 2003). If we look back at the early days of SLA research in the 1970s, it appears that the field did not place much prominence on speaking ability per se and speaking

was predominantly viewed as one of the modes that reflected learners' underlying *interlanguage* system (e.g. Corder, 1971). As such, there was little research that shed light on the nature or acquisition of speaking ability and thus LT researchers had to rely primarily on intuitive expert judgement – rather than empirical evidence – for speaking scale construction.

However, developments in psycholinguistic research provided SLA and LT researchers with a useful framework to understand the nature of speaking. Levelt's (1989) cognitive processing model for speaking remains, to this day, one of the most influential models of L1 speech production, serving as a basis for subsequent L2 speaking models developed in both SLA and LT (e.g. de Bot, 1992; Field, 2011). The wide acceptance of Levelt's model provided a shared theoretical background and helped strengthen the relationship between the two fields.

The growth of speaking research in SLA was tremendous in the 1990s. Following Skehan's (1989) L2 proficiency framework, SLA researchers devised analytical measures for Complexity, Accuracy, and Fluency (CAF) that can usefully quantify L2 speaking performance and reliably predict L2 speaking proficiency (e.g. Révész, Ekiert & Torgersen, 2016; Skehan, 2009; see Chapter 19 in this Volume). In the field of LT, the 1990s saw various new approaches to inform rating scale construction and validation, one of which was an empirical analysis of candidates' speech output (e.g. Fulcher, 1996). The development of CAF measures in SLA was timely, contributing to systematic micro-analyses of candidates' speech samples elicited in speaking tests (e.g. Brown, 2006; Iwashita, Brown, McNamara & O'Hagan, 2008; Tavakoli, Nakatsuhara & Hunter, 2017), and in turn enhancing our understanding of the speaking construct.

Another body of literature to which both SLA and LT research mutually contributed is the examination of the componential structures of L2 speaking proficiency. This is a key line of research for further understanding of the speaking construct and involves an analysis of the extent to which and in what ways holistic, global ratings of speaking proficiency are related to subjective or objective measures of analytical features of speech (e.g. de Jong, Steinel, Florijn, Schoonen & Hulstijn, 2012; Iwashita et al., 2008; McNamara, 1990). To date, probably the most comprehensive study of this kind is de Jong et al.'s (2012) Structural Equation Modelling study, which provided evidence for a multidimensional view of L2 oral proficiency consisting of linguistic knowledge (vocabulary and grammar), processing skills (lexical retrieval and sentence building), as well as pronunciation skills (speech sounds, word stress, and intonation). The significance of linguistic knowledge and processing skills, which explained over three-quarters of the variance in communicative success in speaking in their study, also supports a prominent place for those components in a L2 proficiency model and sheds new light upon earlier models of L2 proficiency, such as Canale and Swain's (1980) and Bachman and Palmer's (1996).

At present, the speaking construct measured in tests is becoming more diversified than ever. On the one hand, expansion of the construct is observed in interactive speaking tests tapping into *interactional competence* (e.g. Galaczi & Taylor, 2018; see Chapter 32 in this Volume), and in interactive and integrated oral test tasks tapping into interactive listening (e.g. Ockey & Wagner, 2018). In these tests, speaking is seen both as a cognitive and a social, interactional trait, resonating Long's (1996) Interactional Hypothesis. On the other hand, we see a narrowing of the speaking construct in semi-direct and automated speaking assessments that are essentially underpinned by a psycholinguistic construct, tapping into *processing competence* (Van Moere, 2012), i.e. the efficiency with which learners process language. These contrasting views to the speaking construct will be revisited throughout this chapter; for example, in terms of how they are represented in test methods or task design. We will also touch on the ways in which technology-mediated speaking tests have started to offer solutions to cover both ends of the construct definition continuum. Here, it is important to note that we do not find the diversification of the measured construct in LT problematic. Our view is aligned with that of Van Moere (2012: 340); "a complementary approach to communicative and psycholinguistic testing will undoubtedly lead to stronger and fairer assessments", and it can offer learners and test users a wide range of choices to select the spoken test that best fits their purposes.

Key issues

Speaking tasks

Research that involves speaking skills requires using speech elicitation tasks. In both LT and SLA, various types of monologic and dialogic/group oral tasks have been used based on evolving theories, views of the speaking construct (Block, 2003), and practical demands. While the focus and purpose of research and tests may be different, there are overlaps in the task types used in the two fields of study (de Jong, 2018) and they share similar rationales for employing one task type over another.

Monologic tasks

Monologic tasks in SLA seem to have been primarily used for studying learners' language competence and cognitive processing. While mechanical tasks such as read-aloud and sentence repetition have been used to investigate learners' interlanguage grammar and degrees of automaticity in processing language (Mackey & Gass, 2005), picture description tasks (with a single picture, a few contrastive pictures, or those in a sequence) that elicit an extended monologue have frequently been used in task-based language teaching (TBLT) research. TBLT aims to examine the effects of manipulating task complexity and task conditions on learner output in order to justify and assist in the pedagogic use of tasks by determining how certain task characteristics affect L2 performance. The two major theories behind TBLT are the Cognition Hypothesis (Robinson, 2001) and the Trade-off Hypothesis (Skehan, 1996), both of which assume that tasks with certain characteristics will impose varying processing loads, which may then direct the attention of L2 learners to different aspects of language use (see Chapters 19 and 30 in this Volume).

In LT, among monologic tasks, mechanical tasks such as read-aloud and sentence repetition can be found in semi-direct or fully automated tests (e.g. PTE Academic, Versant). Echoing the rationale for using such mechanical tasks in SLA, these tasks tap into the processing speed and short-term memory capacity of learners, and are considered indicative of learners' facility with the language (Van Moere, 2012) rather than their ability to deal with the direct interactional demands of face-to-face exchanges. These tasks, while limiting interpretations of test scores, lend themselves particularly well to automated testing approaches, which is discussed in more detail below, as they can increase the robustness of pronunciation and fluency scoring algorithms in automated systems (e.g. Xi, Higgins, Zechner & Williamson, 2012).

Picture description tasks that elicit extended monologues are also common in LT, although their focus is more general than in SLA; they are not so much for finding out about learners' processing and its effects on L2 performance, but for making an evaluative judgement on their speaking proficiency by assessing the ability to describe, compare/contrast, or discuss the picture(s) using the language as expected by the test designers in terms of the lexis, grammar, speed, etc.

Dialogic/group tasks

In SLA, following the rise of the Interaction Hypothesis (Long, 1996) and the Output Hypothesis (Swain, 1985), dialogic/group tasks have been used for interaction-based research (Mackey & Gass, 2005). This type of tasks features information-gaps between the paired/grouped learners, and examples include role play, picture comparison and sequencing, instruction and story-retelling. Using such tasks, SLA researchers have explored the relationship between the types of input, interaction and feedback, as well as the effects of these factors on learning; for example, the effects of using recasts on learning particular morphosyntactic features as compared to other methods of error correction (e.g. Sato & Loewen, 2018).

Similarly, in LT research and practice, dialogic/group oral tasks are used to tap into learners' interactional competence (see Chapter 32 in this Volume). For instance, all Cambridge General English exams now include a paired task where two test-takers have a discussion based on a visual prompt. While a range of LT tasks are available, such as information-gap, ranking, free discussion, and discussion based on a reading text, the focus of LT tasks is again more general than that in SLA. By using these tasks, language testers seek to collect evidence to evaluate how well learners can use various types of language functions (e.g. informational, interactional and interaction management functions;

O'Sullivan, Weir & Saville, 2002) and how effectively they can communicate with others using their interactional resources, rather than to analyse the extent to which certain morphosyntactic features have been acquired.

While the implementation of dialogic/group oral tests is generally considered resource-intensive, thanks to recent advances in technology, we have now started to see the emergence of another strand of research addressing the practical challenges of conducting face-to-face speaking tests while emphasising the importance of including interactional competence as part of the speaking construct. For example, Nakatsuhara, Inoue, Berry, and Galaczi (2017a; 2017b) and Berry, Nakatsuhara, Inoue, and Galaczi (2018) explored the use of video-conferencing technology in delivering the IELTS Speaking Test as a solution for addressing the practical and resource-heavy challenges of conducting face-to-face speaking tests. Also, Ockey, Gu, and Keehner (2017) examined the potential of web-based virtual environments (VEs) as platforms for the delivery and assessment of speaking and communication in real time. The option to simulate real-world environments (e.g. a library or a university) within the assessment setting and immersing learners within these settings are viewed as one of the strengths of VEs in visually reflecting the target language use domain. Such a solution not only encourages interaction but has great potential for authentic task-based language learning bringing together SLA and LT.

Scoring speaking

As well as elicitation tasks, the scoring of the elicited performance is another major factor that contributes to the validity argument in L2 speaking tests. While Chapters 12 and 13 in this Volume discuss rating approaches and issues related to rater behaviour and rater training in depth, here we draw our attention to an inter-disciplinary review of research on fluency; de Jong (2018: 239) notes a lack of objective criteria for measuring fluency in LT as evidenced in the assessment scales of large-scale exams such as IELTS, TOEFL iBT, and ACTFL OPI where the use of vague descriptors may "leave room for subjective interpretation of fluency" by raters. This is in contrast with SLA approaches that predominantly use objectives measures of fluency such as speech rate, mean length of runs in fluency research (de Jong, 2018). The issue of subjective interpretation of assessment scales is also touched on in Isaac's (2018) state-of the-art article on L2 pronunciation.

At the same time, advancements in speech science, automatic speech recognition, and deep neural networking technologies have enabled technology-based forms of testing speaking, and have paved the way for further broadening of the speaking construct by eliciting spontaneous speech and including linguistic features related to lexical use, syntactic complexity, topical coherence, and progression of ideas (Zechner et al., 2015). However, as noted by Chen et al. (2018), in addition to the challenges of speech recognition and associated word error rates, the most problematic areas for an automated system to tackle are those related to the scoring of content, discourse, and topic development. These are higher-level features of speech which are more complex to assess and require a multi-level understanding of speech and communication.

Recommendations for practice

Speaking tasks

We have discussed above that although the rationales for selecting certain task types are largely shared between SLA and LT, the research focus appears narrower and more specific in SLA than in LT. This is because the goals of the two fields of study are different. SLA researchers seek "to understand universal, individual and social forces that influence what gets acquired, how fast, and how well, by different people under different learning circumstances" (Ortega, 2009: 10). Fundamentally, SLA researchers are more interested in investigating variability as evidenced by numerous studies on the effects of individual and contextual factors on L2 learning, such as age, L1, motivation, cognitive styles, types and amount of input, and types and complexity of tasks.

In contrast, language testers are more interested in the stability of L2 measurement and the generalisability of test results (Alderson, 2010). While individual and contextual variables that influence performance and test results are indeed of interest and the effects of some variables such as planning time, interlocutor/rater characteristics and topics have been extensively researched and usefully applied to testing (e.g. Khabbazbashi, 2017; O'Sullivan, 2008;

Wigglesworth & Elder, 2010; Winke, Gass & Myford, 2013), it is neither possible nor fair for tests — at least for large-scale proficiency tests— to cater for all possible sources of variability. The L2 performance has to be elicited through (a set of) the same or parallel tasks and evaluated using the same rating criteria, and the resultant bands of scores or levels need to clearly differentiate among different levels of ability. Otherwise, meaningful and fair comparisons among test-takers cannot be made. Therefore, tasks, rating scales and levels used in LT may appear more general or less individualised than those commonly used in SLA research.

As Alderson (2010) notes, in SLA research, the validity and reliability of the instruments do not seem to have been sufficiently reported. For example, in TBLT, picture-based narrative tasks are frequently used as speech elicitation tools to allow for comparisons of L2 performance across different task manipulations. In order for the results of these studies to be credible, the tasks used in a study must be parallel prior to any manipulation (Inoue, 2013). Otherwise, any differences observed in performances are potentially confounded with inherent differences between the tasks, and counterbalancing the order of task presentation cannot sufficiently address this issue. Although more task-based studies nowadays publish the actual tasks used or reveal the source of where the tasks were obtained (e.g. Foster & Tavakoli, 2009; Yuan & Ellis, 2003), only a few provide empirical research evidence of the parallelness of tasks beforehand (Inoue, 2013; Weir & Wu, 2006).

In LT, it is always recommended that test tasks are developed based on test specifications, which control the complexity of the prompts and expected performance in terms of the levels of vocabulary, grammar, functions, discourse type, etc. (Taylor, 2011; Weir, 2005). Parallel versions of the tasks are created and piloted, so as to ensure the performances from different administrations can be meaningfully compared (see Chapter 43 in this Volume). This practice in LT seems quite relevant to SLA, contributing to enhancing the confidence in and generalisability of valuable findings of SLA speaking research.

Scoring speaking

An alternative suggestion to subjective, rater-based scoring approaches is the use of more objective measures facilitated by technology and automated approaches to assessing speaking (e.g. Bernstein, Van Moere & Cheng, 2010). Notwithstanding that these have their own limitations, they can address some of the concerns raised about subjectivity in assessment and bridge the distance between the fields of SLA and LT. At the same time, while automated assessment solutions can detect hundreds of features in speech, it is important for LT researchers to draw on results of SLA studies to incorporate, in the scoring algorithms, those features which are shown to be valid measures of the construct of interest. For example, while fluency features of speech rate and pausing are commonly used in automated assessment systems (Chen et al., 2018), several studies in SLA have pointed to the importance of the *location* of pauses as exerting more influence on listeners' perceptions of fluency than the frequency of pauses (e.g. Tavakoli, 2010).

Automated scoring systems rely heavily on constrained and monologic task types. A possible negative impact is therefore an excessive preoccupation with monologic speech in learning contexts at the expense of more interactive tasks and co-constructed dialogues. Another critical issue – closely related to the narrow test construct – is the increased likelihood of candidates displaying abnormal test behaviours or applying strategies in an attempt to cheat or fool automated systems. As a promising solution for retaining the wide construct coverage while making the best use of automated scoring systems, we would like to echoe the suggestions by Khabbazbashi and Galaczi (2016), Isaacs (2018), and Lim (2018), and argue for a complementary human-machine approach to assessing speaking with raters focusing on aspects of speech that are too complex for automatic evaluation and machines measuring relevant features of speech that can be automatically derived and which have been shown to be reliable predictors of oral proficiency.

Feedback

Amongst the criticisms levelled at LT researchers in Shohamy (2000) was a lack of sensitivity to learners and candidates in terms of their individual learning needs. Since then, the field of LT has come a long way in increasing recognition of the importance of learning-oriented approaches to assessment (LOA) and promoting learning by creating a synergy

between instruction, assessment, and learning (e.g. Jones & Saville, 2016). Technology has great potential in facilitating this synergy by providing timely and individualised feedback which, according to SLA theories, can promote L2 development (Gass & Mackey, 2015). For example, Franco et al. (2010) designed a toolkit that can score the pronunciation of non-native speakers and provide explicit diagnostic feedback at the phone level for specific pronunciation mistakes; a functionality which can greatly assist teachers in L2 learning contexts where providing individualised support may be practically unfeasible. Morton, Gunson, and Jack (2012) designed a spoken dialogue system where success on the task was defined, not as score targets, but as completing a real-world activity such as buying train tickets through interacting with a series of conversational agents in a web environment. The limitations in automatic speech recognition of this system were countered by the provision of 'implicit' feedback such as recasts, repetitions, reformulations, and hints when the system encountered errors in learners' utterance or where learners failed to provide a response. While technical challenges remain in automating spoken interactions and the provision of feedback (Litman, Strik & Lim, 2018), these examples and the surge of research on educational games, spoken dialogue systems, and whole tutoring systems demonstrate the usefulness of technology in bringing together learning and assessment in innovative ways. Moreover, this integration of learning and assessment and the capacity for systems to collect and track language use data over extended periods of time can greatly advance our understanding of the process of language acquisition and provide exciting avenues for joint research between the fields of SLA and LT.

Testing tips

- When selecting or designing speaking test tasks, always ask: *Where can the construct we wish to measure be located in the cognitive-social construct spectrum of speaking assessments? Are we prioritising, for example, automaticity (speed of processing language), or is our focus more on interactional effectiveness?*
- Your answers to these types of questions will help you choose appropriate speaking task formats and scoring methods, and provide you with a selection of available technologies.
- Technologies in delivering and scoring speaking tests and offering individualized feedback are highly useful, as long as we are fully aware of their pros and cons in relation to the construct we wish to measure.

Recommended Readings

Fulcher, G. (2003). *Testing second language speaking*. London: Longman/Pearson Education.

This book, although written some time ago, provides a thorough guide into designing and implementing speaking tests, as well as useful and critical summaries of research in L2 speaking assessment.

Lim, G. (Ed.). (2018). *Conceptualizing and operationalizing speaking assessment for a new century* [Special issue], *Language Assessment Quarterly*, *15*(3).

The articles in this special issue consider important aspects of the speaking construct such as interactional competence, collocational competence, fluency, and pronunciation and whether and to what extent they have been assessed while reflecting on the potential role of technology in enhancing assessment practices.

Taylor, L. (2011). *Examining speaking: Research and practice in assessing second language speaking*. Cambridge: UCLES/Cambridge University Press.

This edited volume presents a review of relevant literature on the assessment of speaking and provides a systematic framework for validating speaking tests.

References

Alderson, J. C. (2010). Language testing-informed SLA? SLA-informed language testing? In I. Bartning, M. Martin, & I. Vedder (Eds.), *Communicative proficiency and linguistic development: intersections between SLA and language testing research* (pp.239-248). EUROSLA.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Berry, V., Nakatsuhara, F., Inoue, C., & Galaczi, E. (2018). Exploring the use of video-conferencing technology to deliver the IELTS Speaking Test: Phase 3 technical trial. *IELTS Partnership Research Papers, 2018/1*. Retrieved from: www.ielts.org/-/media/research-reports/ielts-research-partner-paper-4.ashx

Bernstein, J., Van Moere, A., & Cheng, J. (2010). Validating automated speaking tests. *Language Testing, 27*(3), 355-377. https://doi.org/10.1177/0265532210364404

Brown, A. (2006). Candidate discourse in the revised IELTS Speaking Test. *IELTS Research Reports, 6,* 71–89. Retrieved from: www.ielts.org/-/media/research-reports/ielts_rr_volume06_report3.ashx

Block, D. (2003). *Social turn in second language acquisition*. Washington, D.C.: Georgetown University Press.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*, 1–47. https://doi.org/10.1093/applin/I.1.1

Chen, L., Zechner, K., Yoon, S., Evanini, K., Wang, X., Loukina, A., . . . Ma, M. (2018). Automated scoring of nonnative speech using the SpeechRater SM v. 5.0 engine. *ETS Research Report Series.* https://doi.org/10.1002/ets2.12198

Corder, S. P. (1971). Idiosyncratic dialects and error analysis. *International Review of Applied Linguistics 9*, 147-159. 10.1515/iral.1971.9.2.147

De Bot, K. (1992) A bilingual production model: Levelt's 'Speaking' model adapted. *Applied Linguistics*, *13*(1),1–24. https://doi.org/10.1093/applin/13.1.1

De Jong, N. (2018). Fluency in second language testing: Insights from different disciplines. *Language Assessment Quarterly, 15*(3), 237-254. https://doi.org/10.1080/15434303.2018.1477780

De Jong, N., Steinel, M., Florijn, A., Schoonen, R., & Hulstijn, J. (2012). Facets of speaking proficiency. *Studies in Second Language Acquisition, 34*(1), 5-34. https://doi.org/10.1017/S0272263111000489

Field, J. (2011). Cognitive validity. In L. Taylor (Ed.), *Examining speaking: Research and practice in assessing second language speaking* (pp. 65-111). Cambridge: UCLES/Cambridge University Press.

Foster, P. & Tavakoli, P. (2009). Native speakers and task performance: Comparing effects on complexity, fluency and lexical diversity. *Language Learning, 59*(4), 866–896. https://doi.org/10.1111/j.1467-9922.2009.00528.x

Franco, H., Bratt, H., Rossier, R. Rao Gadde, V., Shriberg, E., Abrash, V., & Precoda, K. (2010). EduSpeak: A speech recognition and pronunciation scoring toolkit for computer-aided language learning applications. *Language Testing, 27*(3), 401-418. https://doi.org/10.1177/0265532210364408

Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction, *Language Testing, 13*(2), 208–238. https://doi.org/10.1177/026553229601300205

Fulcher, G. (2003) *Testing second language speaking.* London: Longman.

Galaczi, E., & Taylor, L. (2018). Interactional competence: Conceptualisations, operationalisations, and outstanding questions. *Language Assessment Quarterly, 15*(3), 219-236. https://doi.org/10.1080/15434303.2018.1453816.

Gass, S. M., & Mackey, A. (2015). Input, interaction, and output in second language acquisition. *Theories in second language acquisition: An introduction* (Second Edition ed., pp. 180-206). New York: Routledge.

Hulstijn, J. H. (2011). Language proficiency in native and nonnative speakers: An agenda for research and suggestions for second-language assessment. *Language Assessment Quarterly, 8*(3), 229-249. https://doi.org/10.1080/15434303.2011.565844

Inoue, C. (2013). *Task equivalence in speaking tests: Investigating the difficulty of two spoken narrative tasks*. Bern: Peter Lang.

Issacs, T. (2018). Shifting sands in second language pronunciation teaching and assessment research and practice. *Language Assessment Quarterly, 15*(3), 273–293. https://doi.org/10.1080/15434303.2018.1472264

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics, 29*(1), 24-29. https://doi.org/10.1093/applin/amm017

Jones, N., & Saville, N. (2016). *Learning oriented assessment: A systemic approach.* Cambridge: UCLES/Cambridge University Press.

Khabbazbashi, N. (2017). Topic and background knowledge effects on performance in speaking assessment. *Language Testing, 34*(1), 23-48. https://doi.org/10.1177/0265532215595666

Khabbazbashi, N., & Galaczi, E. (2016). *Technology and the 'what' of automated speaking tests*. Paper presented at the pre-conference SIG meeting of the 13th EALTA Conference, Valencia, Spain.

Levelt, W. J. M. (1989). *Speaking.* Cambridge, MA: Mit Press.

Lim, G. S. (2018). Conceptualizing and operationalizing second language speaking assessment: Updating the construct for a new century, *Language Assessment Quarterly, 15*(3), 215-218. https://doi.org/10.1080/15434303.2018.1482493

Litman, D., Strik, H., & Lim, G. S. (2018) Speech technologies and the assessment of second language speaking: Approaches, challenges, and opportunities, *Language Assessment Quarterly, 15*(3), 294-309. https://doi.org/10.1080/15434303.2018.1472265

Long, M. H. (1996). The role of the linguistic environment in second language acquisition. In W. Ritchie and T. Bhatia (Eds.), *Handbook of second language acquisition* (pp.413-468). New York: Academic Press.

Mackey, A., & Gass, S. M. (2005). *Second language research: Methodology and design*. Mahwah, NJ: Lawrence Erlbaum Associates.

McNamara, T. F. (1990). Item response theory and the validation of an ESP test for health professionals. *Language Testing, 7*(1)*,* 52–76. https://doi.org/10.1177/026553229000700105

Morton, H., Gunson, N., & Jack, M. (2012). Interactive language learning through speech-enabled virtual scenarios. *Advances in Human-Computer Interaction, 2012*, 1-14. https://doi.org/10.1155/2012/389523

Nakatsuhara, F., Inoue, C., Berry, V., & Galaczi, E. (2017a). Exploring the use of video-conferencing technology in the assessment of spoken language: A mixed-methods study. *Language Assessment Quarterly, 14*(1), 1-18. https://doi.org/10.1080/15434303.2016.1263637

Nakatsuhara, F., Inoue, C., Berry, V., & Galaczi, E. (2017b). Exploring performance across two delivery modes for the IELTS Speaking Test: Face-to-face and video-conferencing delivery (Phase 2*). IELTS Research Reports Online Series, 2017/3*, 1-74. Retrieved from: www.ielts.org/-/media/research-reports/ielts-research-partner-paper-3.ashx

Ockey, G. J., Gu, L., & Keehner, M. (2017). Web-based virtual environments for facilitating assessment of L2 oral communication ability. *Language Assessment Quarterly, 14*(4), 346-359. https://doi.org/10.1080/15434303.2017.1400036

Ockey, G., & Wagner, E. (2018). An overview of interactive listening as part of the construct of interactive and integrated oral test tasks. In G. Ockey, & E. Wagner (Eds.), *Assessing L2 listening: Moving towards authenticity* (179-192). Amsterdam, Philadelphia: John Benjamins.

Ortega, L. (2009). *Understanding second language acquisition*. London, NY: Routledge.

O'Sullivan, B. (2008). *Modelling performance in tests of spoken language.* Frankfurt: Peter Lang.

O'Sullivan, B., Weir, C.J., & Saville, N. (2002). Using observation checklists to validate speaking-test tasks. *Language Testing, 19*(1), 33-56. https://doi.org/10.1191/0265532202lt219oa

Palmer, H. E. (1921). *The oral method of teaching languages.* Cambridge: Heffers.

Révész, A., Ekiert, M., & Torgersen, E. (2016). The effects of complexity, accuracy, and fluency on communicative adequacy in oral task performance. *Applied Linguistics, 37*(6), 828-848. https://doi.org/10.1093/applin/amu069

Robinson, P. (2001). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics 22*(1), 27-57. https://doi.org/10.1093/applin/22.1.27

Sato, M. & Loewen, S. Metacognitive instruction enhances the effectiveness of corrective feedback: Variable effects of feedback types and linguistic targets. *Language Learning, 68*(2), 507-545. https://doi.org/10.1111/lang.12283

Shohamy, E. (2000). The relationship between language testing and second language acquisition, revisited. *System, 28*(4), 541-533. https://doi.org/10.1016/S0346-251X(00)00037-3

Skehan, P. (1989). *Individual differences in second language learning.* London: Edward Arnold.

Skehan, P. (1996). A framework for the implementation of task based instruction. *Applied Linguistics, 17*(1), 38-62. https://doi.org/10.1093/applin/17.1.38

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency and lexis. *Applied Linguistics, 30*(4), 510-532. https://doi.org/10.1093/applin/amp047

Swain, M. (1985). Communicative competence: some roles of comprehensible input and comprehensible output in its development. In S. M. Gass and C. G. Madden (Eds.), *Input in second language acquisition* (pp.235-253). Rowley, MA: Newbury House.

Sweet, H. (1899). *The practical study of languages.* London: Dent.

Tavakoli, P. (2011). Pausing patterns: Differences between L2 learners and native speakers. *ELT Journal, 65*(1), 71–79. https://doi.org/10.1093/elt/ccq020

Tavakoli, P., Nakatsuhara, F., & Hunter, A. M. (2017) Scoring validity of the Aptis speaking test: Investigating fluency across tasks and levels of proficiency. *ARAGs Research Reports Online, AR-G/2017/7*, 1-56. Retrieved from: www.britishcouncil.org/sites/default/files/tavakoli_et_al_layout.pdf

Taylor, L. (2011). *Examining speaking: Research and practice in assessing second language speaking.* Cambridge: UCLES/Cambridge University Press.

Van Moere, A. (2012). A psycholinguistic approach to oral language assessment. *Language Testing, 29*(3), 325-344. https://doi.org/10.1177/0265532211424478

Vidaković, I., & Galaczi, E. G. (2013). The measurement of speaking ability 1913-2012. In C. J. Weir, I. Vidaković, & E. D. Galaczi (Eds.). *Measured constructs: A history of Cambridge English Language Examinations 1913–2012* (pp.257-346). Cambridge: UCLES/Cambridge University Press.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach.* Basingstoke: Palgrave Macmillan.

Weir, C. J., & Wu, J. R. W. (2006). Establishing test form and individual task comparability: A case study of a semi-direct speaking test. *Language Testing, 23*(2), 167–197. https://doi.org/10.1191/0265532206lt326oa

Wigglesworth, G., & Elder, C. (2010). An investigation of the effectiveness and validity of planning time in speaking test tasks. *Language Assessment Quarterly, 7*(1), 1-24. https://doi.org/10.1080/15434300903031779

Winke, P., Gass, S., & Myford, C. (2013). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing, 30*(2), 231-252. https://doi.org/10.1177/0265532212456968

Xi, X., Higgins, D., Zechner, K., & Williamson, D. (2012). A comparison of two scoring methods for an automated speech scoring system. *Language Testing, 29*(3), 371-394. https://doi.org/10.1177/0265532211425673

Yuan, F. & Ellis, R. (2003). The effects of pre-task planning and online planning on fluency, complexity and accuracy in L2 monologic oral production, *Applied Linguistics, 24*(1), 1–27. https://doi.org/10.1016/j.sbspro.2011.11.445

Zechner, K., Chen, L., Davis, L., Evanini, K., Lee, C. M., Leong, C. W., . . . Yoon, S. (2015). Automated scoring of speaking tasks in the test of English-for-Teaching (TEFT™). *ETS Research Report Series, 2015*(2), 1-17. https://doi.org/10.1002/ets2.12080